International Journal of Advanced Computer and Mathematical Sciences
ISSN 2230-9624. Vol 4, Issue 2, 2013, pp136-142
http://bipublication.com

OPTIMALITY CRITERIA FOR CLASSIFICATION THROUGH ROC CURVE ANALYSIS IN THE PRESENCE OF OUTLIERS

R V Vardhan[1, 2], Sarma KVS[2], Alladi Mohan[3] and Suchitra M M[4]

[1]Department of Statistics, Pondicherry University, Puducherry-14
[2]Department of Statistics, Sri Venkateswara University, Tirupati, A.P, INDIA.
[3]Department of Medicine, Sri Venkateswara Institute of Medical Sciences, Tirupati – 517502
[4]Department of Biochemistry, Sri Venkateswara Institute of Medical Sciences, Tirupati – 517502

[Received-28/09/2012, Accepted-11/03/2013]

ABSTRACT

In this paper, we focus on fitting non-parametric and Binormal ROC Curves using spreadsheet functions and illustrate the procedures by using TB data collected in a hospital study. The Receiver Operating Characteristic (ROC) curves is one of those statistical techniques that are now ubiquitous in a wide variety of substantive fields and it is a well accepted measure of accuracy for tests with both continuous and ordinal results. Tuberculosis (TB) is a major health problem often called "the captain of these men of death". Pleural Fluid Adenosine Deaminase (ADA) is a known biomarker to determine the presence to TB as against the gold standard called IFN-Gamma which is both costly and not affordable to many patients. The optimal cutoff value which provides a better accuracy of classification is determined. The effect of age and sex as cofactors is also considered for classification of patients using linear discriminant function (LDF). In this paper, we have studied the discriminating ability of pleural fluid ADA in the diagnosis of TB. The ROC analysis has shown that a cut-off value of 36 IU/L provides the optimal trade-off between sensitivity and specificity. It is observed that the age and sex of the subject may influence the decision criterion.

Keywords: ROC Curve Analysis, Linear Discriminant Function, Area under the Curve, Adenisine Deaminase (ADA)

INTRODUCTION

ROC curve analysis gives the accuracy of a diagnostic test. It measures the test ability to discriminate among alternative states of health. A generally accepted rule that correctly classifies diseased and non-diseased cases often exists which is known as *gold standard* for classification. The discriminating ability of a classifier is often compared with the classification made by the gold standard using one or more measures of performance. Much work in the area of ROC curves was reported by Green and Swets (1966) and Swets et.al (1979) [1, 2]. Leo Lusted, a Radiologist suggested using ROC analysis in Medical decision making in 1967 and his original description of ROC is, "Plotting false - positive diagnosis and true - positive diagnosis on the X - axis and Y - axis respectively". Metz (1978) stated that ROC analysis is useful to determine the discriminating ability of a diagnostic test [3]. In later years, eventually ROC analysis made its way into other areas of medicine. ROC curve analysis gives the accuracy of a diagnostic test. Diagnostic accuracy is the most fundamental characteristic of the test itself as a classification device. It measures the test ability to discriminate among alternative states of health. Let X denotes the test result (which is continuous or discrete random variable) and c denotes a cut-off value or threshold. Assume that a subject is classified as positive if X > c and negative otherwise (this rule could be the other way also). Each subject will be in one of the disjoint states D and H denoting the Diseased and Healthy states respectively. From the available data we get the following cases.

1. True Positive (TP): Number of cases where both diagnosis and test are Positive.
   $$TP = [X \geq c | D]$$
2. False Positive (FP): Number of cases where the diagnosis is negative but the test is positive. $FP = [X \geq c| H]$
3. True Negative (TN): Number of cases where both the diagnosis and test is Negative.
   $$TN = [X > c| D]$$
4. False Negative (FN): Number of cases where the diagnosis is positive but the test is negative. $FN = [X < c| H]$
5. *Sensitivity*(S_n) or True Positive fraction (TPF) $= TP\ (TP + FN)^{-1} = P\ [X > c| D]$ and
6. *Specificity* (S_p)or False Positive fraction (FPF) $= FP\ (FP + TN)^{-1}= P\ [X>c|H]$

Both S_n and S_p takes values between 0 and 1. Further if the value of c increases, both S_n and S_p will decrease. A good diagnostic test is supposed to have high sensitivity with corresponding reasonably low specificity. The ROC curve is a plot of S_n values against (1-S_p) values and the diagnostic effciency is often judged by the shape of the ROC curve. Each cut-off C_i corresponds to a point (1-S_p, S_n) on a ROC curve. The graph of ROC Curve is of unit square on X (1-specificity) and Y (Sensitivity) axis. The value of ROC lies between 0 and 1 (Krzanowski and Hand, 2009)[4]. If the test is perfect then $1\text{-}S_p = S_n= 1$, and if $1\text{-}S_p = S_n\ (\neq 1)$ the ROC curve becomes a straight line from (0, 0) to (1, 1).

The accuracy of a diagnostic test can be expressed in terms of the Area Under the Curve (AUC). It measures the ability of the test to discriminate between diseased and non-diseased. If X and Y denote the test result in D and H groups respectively, then AUC is equivalent to P[X > Y] [5]. Higher the AUC, better will be the diagnostic test. Value of AUC = 0.5 indicates 50% chance for correct classification and a good test shall have AUC > 0.5. The diagonal line connecting the origin with the point (1,1) is the line of indifference and a good ROC curve shall be far away to the left of this line. Tests with AUC < 0.5 are considered to be worst cases.

There are several methods of calculating the AUC and the Trapezoidal rule is one popular one, known as non-parametric method. Binormal model is a different but a parametric approach to carry out the ROC analysis. It is considered better than the non-parametric method and works well when the data in both the D and H groups follows normal distribution.

In this paper we focus on fitting both the non-parametric and binormal ROC curve model in the presence of outliers using plueral fluid ADA data. Optimal cutoff value for ADA has been determined. The effect of age and sex as cofactors has also been considered in this paper.

ROC Curve Estimation - Non-parametric ROC Curve

The test score of a healthy patient is represented as a real random variable X with distribution function F and density f. Similarly a diseased patients score will be denoted by Y with distribution function G and density g; X and Y are independent. The sensitivity of the test is defined as $S_n(c) = 1\text{-}\ G(c)$, which is the probability of correctly classifying a diseased individual when cutoff point c is used. Similarly we define the tests specificity $S_p(c) = F\ (c)$. as the probability of correctly classifying a healthy patient. Clearly these are the complements of the familiar Type I and Type II errors. The receiver operating characteristic curve (ROC) is defined as a plot of the “true positive fraction”, S_n(c), on the vertical axis versus the “false positive fraction” 1-SP(c), on the horizontal axis as c varies from $- \infty$ *to* $+\infty$ (Hsieh and Turnbull ,1992)[6] . Equivalently, ROC curve can be viewed as

$ROC(t) = 1 - G(F^{-1}(1 - t));\ 0 \leq t \leq 1$ (1)

Area under the Curve: Bamber (1975)[5] noted that the area under the empirical ROC curve is equal to the Mann-Whitney U-statistic and is given

$$U = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left[I\left(S_{Pi} > S_{Nj}\right) + \frac{1}{2} I\left(S_{Pi} = S_{Nj}\right) \right] \qquad (2)$$

The formal definition of the Area under the curve (AUC) Krzanowski and Hand (2009) [4] is given by

$$AUC = \int_0^1 y(x)\,dx = P\left(S_n > S_p\right) \qquad (3)$$

Since the data was found to follow normal distribution, the Binormal model is expected a produce better results than the empirical (non-parametric) ROC. The following section deals with this aspect.

Binormal ROC Curve Model

Let d_i and h_j are the test values in the D and H groups and these are normally distributed.

$d_i \sim N(\mu_D, \sigma_D^2)$ and $h_j \sim N(\mu_H, \sigma_H^2)$. It is also assumed that there exists a monotonic function that transforms these values into normal distributions if the values are not normal.

Then the ROC model is given by

$SN = \Phi(a + b\Phi^{-1}(1 - SP));\ 0 \leq SP \leq 1$ (4)

where a and b are constants and Φ(.) is the standard normal distribution.

The parameters a and b are estimated as $a = \frac{\mu_D - \mu_H}{\sigma_H}$ and $b = \frac{\sigma_D}{\sigma_H}$ and the AUC is given by

$$AUC = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) \qquad (5)$$

Farragi and Reiser (2002) [7] have conducted extensive simulations and demonstrated that the binormal model performs very similar to the non parametric approach. However, when the underlying distributions are poorly separated (get overlapped) then binormal model may not be satisfactory.

A Case Study in detection of Tuberculosis

The data were collected from a tertiary care teaching hospital, Sri Venkateswara Institute of Medical Sciences (SVIMS), Tirupati during 2004-2008. One hundred patients (66 males) were included in the study. The data on each subject (patient) included the parameters age, sex, diagnosis Status (non-TB = 0 and TB = 1) and pleural fluid ADA level (IU/L). Tuberculosis (TB) is major health problem globally and has been called "the captain of these men of death". In diagnosing TB pleural effusion, two well known methods are Adenosine Deaminase (ADA) and interferon -γ (IFN- γ) (Piras et. al, 1978)[8]. Of these two, ADA estimation has attracted the attention of clinical researchers in recent times. The studies on usefulness of ADA estimation in the diagnosis of TB pleural effusion has been carried out by (Sharma et. al, 2001, 2006)[9] and (Cimen et. al, 2008)[10]. Using the threshold value of ADA 35 IU/L (Sharma. et. al, 2001), the pleural effusion have been classified into TB pleural effusion and Non-TB pleural effusion. After classification, it is noticed that (n_1 =) 67 (67%) had TB pleural effusion and (n_2 =)33 (33%) had non-TB pleural effusion. Out of 67 diseased subjects, 21 (31.3 %) were females. Similarly, out of 33 healthy subjects, 13 (39.4%) were females. The age group covered in the study is from 16 to 75 years.

Test for normality of data - Detection of Outliers

Hypothesis of data normality has been verified using P-P plot and also Kolmogorov-Smirnov (K-S) test for the values in each group. The p-p plot for pleural fluid ADA levels in subjects with non-TB pleural effusion has shown good normal fit with mean of 20.48 IU/L, standard deviation of 8.519 IU/L, though all the plotted points are not close to the diagonal line. K-S test has given Z = 0.824 (p = 0.506) and hence normality groups is established. For the TB group (Figure 2.1(b)) normality was also established (mean = 76.27 IU/L and SD= 26.022 IU/L, Z = 0.669, p = 0.763 for the K-S test).

The presence of outliers, if any, was examined and the 5% outliers at the top and bottom are examined for possible removal in the analysis.

These values are filtered out and the trimmed mean was found to be 20.61 IU/L in the healthy group and 74.51 IU/L in the diseased group. This mean is a more reliability estimate of the average pleural fluid ADA instead of the mean in the presence of outliers. It is seen that the outliers are far away from the trimmed mean. We first retain the outliers and obtain the results and later compare the results with outliers removed.

The results are found to be different when the outliers are removed as shown in table 1 When the outliers are removed the data set is left out with n_1 = 23 cases in H-group and n_2 = 57 cases in D-group. The following results provide a comparison between the two data sets, with outliers and without outliers. It can be observed that removal of the outliers has resulted in higher AUC in case of Binormal ROC, while there is no change in the nonparametric AUC. The Binormal model had a better fit when outliers are removed. By removing the outliers the mean distance between populations is better identified when compared to the case with outliers.

Since pleural fluid ADA values are continuous each value was considered as a possible threshold value and the S_n and S_p values were calculated by designing an Excel template and utilizing the paste function COUNT and COUNTIF. Figure-2 shows the plotted ROC curve.

The AUC basing on the trapezoidal rule could be found by the following Excel function.

AUC= (SUMPRODUCT(F2:F100,E3:E101)-SUMPRODUCT(E2:E100,F3:F101)+E101*F101-E2*F2)/2

The optimal cut-off was found to be 36 IU/L with high sensitivity (Sn = 1) and specificity (Sp = 1) and AUC = 1.00, which means, by using this cut-off one can classify the healthy as healthy and diseased as diseased with 100% accuracy with pleural fluid ADA as the biomarker. Earlier S.K.Sharma et. al (2001) obtained the appropriate cut-off as 35 IU/L with 83.3 % sensitivity and 66.6% specificity. The empirical ROC curve is shown in Figure 2

The fitted ROC model is found to be ROC = $\Phi(z)$ where z = 6.5483 + 3.0546Φ^{-1}(1-Sp). The linear function z is a measure of discriminating ability of the threshold or cut-off. The intercept estimates the standardized gap between the means of the two groups while the slope estimates the scale with which z increases. The AUC is estimated as 0.9792. This is lower than the AUC found from the non-parametric method which gave AUC = 1.00. One possible reason for this is that the data in each group contained some extreme values which have influenced the parameters a and b of the Binormal model. The plotted Binormal ROC curve is shown in Figure – 3.

The Binormal ROC curves with and without outliers are shown in Figure – 4

It can be seen that, when the outliers are removed the ROC curve shows better shape. In the following section we consider a Linear Discriminant Function (LDF) to accommodate the cofactors like age and sex in the classification of subject using pleural fluid ADA.

The Effect of Age and Sex as Cofactors

The effect of cofactors like age and sex along with ADA can be brought into the frame work of LDF so that a new subject with unknown group membership can be classified into TB / Non-TB group with a reasonable level of accuracy.

The model used is

$U = a_o + a_1 * ADA + a_2 * AGE + a_3 * SEX + \varepsilon$

where U is the discriminant score and ε is the error component. Using the data after removing the outliers, the following LDF is obtained with stepwise regression. The Discriminant function itself acts as a bio marker for classification. If the estimated score U for new subject exceeds a cut-off, it is classified as TB and non-TB otherwise.

1. The LDF is found to be

$U = -2.935 + 0.063 * ADA -0.020 * AGE$

(SEX is not included in the model (F= 2.312, p-value= 0.926))

2. Function value at the group centroids and the cut-off is -0.753.

Group	Function Score at the Centroids
TB Group	1.019
Non-TB group	-2.525
Average (Cut-o®)	**-0.753**

The percentage of correct classification is 98.8%

3. The ROC curve for this LDF is shown in Figure-5

The AUC of this LDF is 1.00 which means 100% correct classification is done. It can be seen that the stepwise LDF offers a better classification of subjects into correct groups using pleural ADA and AGE as the determinants of TB status.

This result helps in the screening of patients with unknown TB status. We have also run two other regression models to generate the LDF. The following results are obtained. It is also observed that the percentage of correct classification has increased with different discriminant functions.

RESULTS AND DISCUSSION

In this paper, we have studied the discriminating ability of pleural fluid ADA in the diagnosis of TB. The data collected at SVIMS hospital has been used to identify the optimal pleural fluid ADA cut-off level as a marker. The ROC analysis has shown that a cut-off value of 36 IU/L provides the optimal trade-off between sensitivity and specificity. It is observed that the age and sex of the subject may influence the decision criterion. Hence, a Linear Discriminant Function has been developed with age as a cofactor. It is found that this LDF provides an optimal classification with 98.8% accuracy of classification.

REFERENCES

1. Green, D.M and Swets, J.A , Signal Detection theory and Psychophysics , Wiley, Newyork.
2. Swets et. al (1979), Assessment of Diagnostic Technologies ,*Science*, **205**,753 -759
3. Metz CE. **(***1978)* Basic Principles of ROC analysis , *Seminars in Nuclear Medicine* **8**, 283-298.
4. Krzanowski W.J and Hand D.J (2009), ROC Curves for Continuous Data, , CRC Press,London.
5. Bamber D. **(**1975) The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* **12**, 387-415.
6. Hsieh and Turnbull **(**1996), Nonparamteric and Semiparametrc Estimation of the Receiver Operating Characteristic Curve, *Annals of Statistics* **24**, 1**:**25-40.
7. Farragi D, Reiser B **(**2002). Estimation of the area under the ROC curve , *Statistic in Medicine* **21**, 3093 -3106
8. Piras et.al (1978), Adenosine deaminase activity in pleural effucions: An aid to differential diagnosis , *British Medical Journal* **2**, 1751 - 1752.
9. S. K. Sharma. et.al (2006), Diagnostic Accuracy of Ascitic Fluid IFN - γ and Adenosine Deaminase Assays in the Diagnosis of Tuberculosis Ascities, *Journal of Interferon and Cytokine research* **26**, 484-488.
10. imen et. al (2008), The Relationship Between Serum Adenosine Deami-nase Levels in Lung Tuberculosis Along with Drug Resistance and the Category of Tuberculosis , *Turkish Respiratory Journal* **9(1)**, 20-23.
11. S. K. Sharma. et.al (2001) , A Prospective Study of Sensitivity and Specificity of Adenosine Deaminase Estimation in the Diagnosis of Tuberculosis Pleural Effusion , *Indian Journal of Chest Dis Allied Science* **43**, 149-155.

S. No	Characteristics	With Outliers	Without Outliers
1	n_1	33	23
2	n_2	67	57

3	- K-S Test: H Group
	- K-S Test: D Group
		Z = 0.824;
		p-value = 0.506
		Z = 0.669;
		p-value = 0.763
			Z = 0.694;
			p-value = 0.721
			Z = 0.601;
			p-value = 0.863

4	Trimmed Mean	H-Group: 20.61 D-Group: 74.51	H-Group: 20.56 D-Group: 73.97
5	- AUC : Nonparametric - AUC: Binormal	1.00 0.979	1.00 0.997
6	- Mean ± S.D for H Group - Mean ± S.D for D	20.48 ± 8.519 76.27 ± 26.022	20.65 ± 6.678 74.07 ± 18.086

	Group		
7	Optimal Cut-off	36 IU/L	36 IU/L
8	S_n	1.00 (100%)	1.00 (100%)
9	S_p	1.00 (100%)	1.00 (100%)
10	Binomial Parameters	a = 6.5483; b = 3.0546	a = 7.9986; b = 2.7081

Table 1: Comparative results when outliers are removed

Method	Model	% Classified
Full Regression with ADA	D = -3.733 + 0.064 * ADA	96.3%
Full regression with ADA, AGE and SEX	D = -3.081 + 0.062*ADA -0.023 * AGE + 0.433 * SEX	97.5%
Stepwise regression with ADA, AGE	D = -2.935 + 0.063* ADA - 0.020 * AGE	98.8%

Table 2

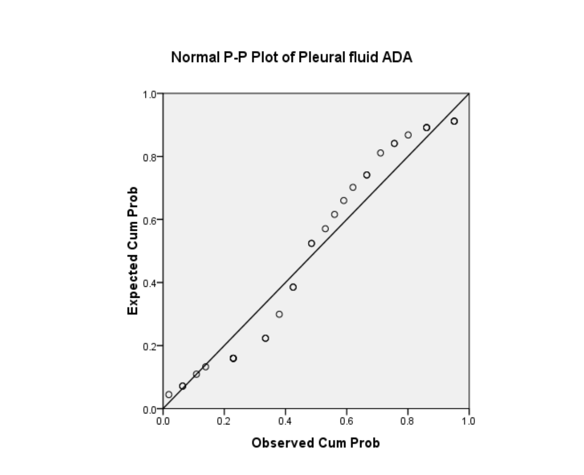

Figure: 1 (a) p-p plot for Healthy subjects

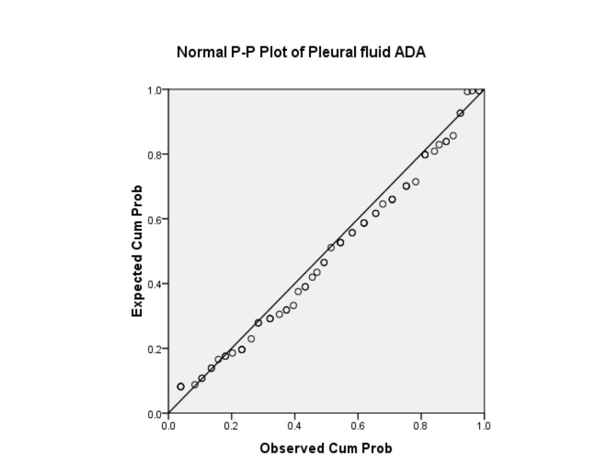

Figure: 1 (b) p-p plot for Diseased subjects

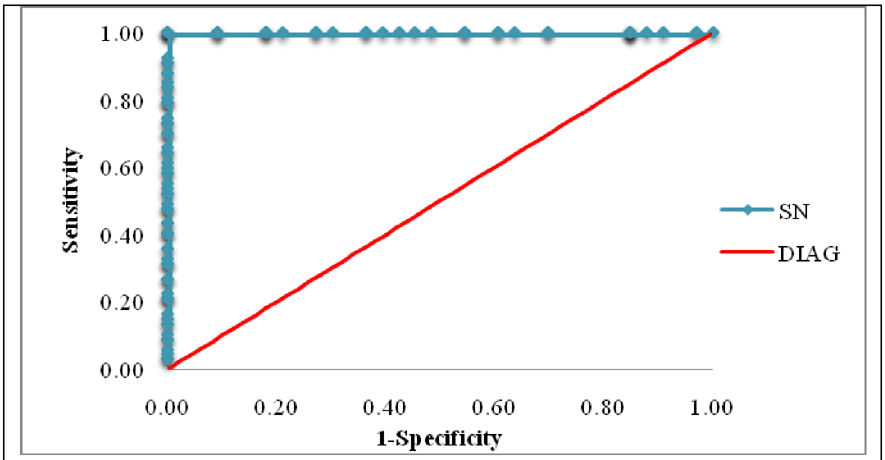

Figure-2: Empirical ROC Curve for ADA Data

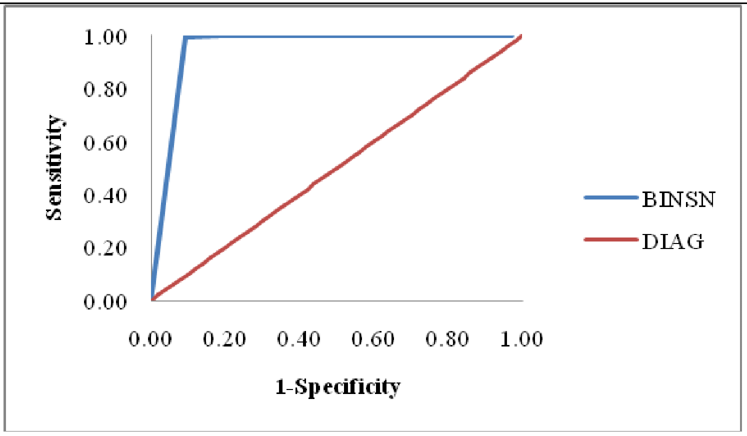

Figure 3: Binormal ROC curve for ADA data

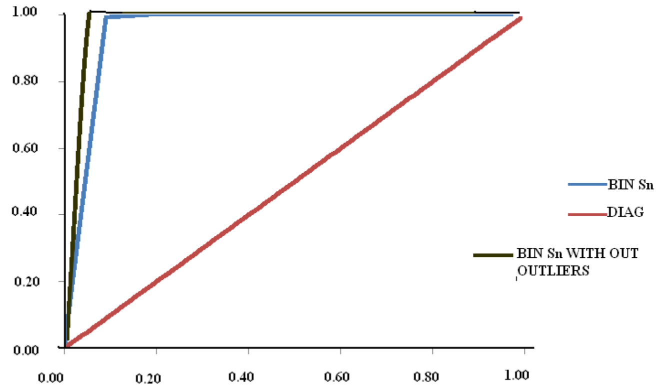

Figure: 4: Binormal ROC Curves with and without outliers

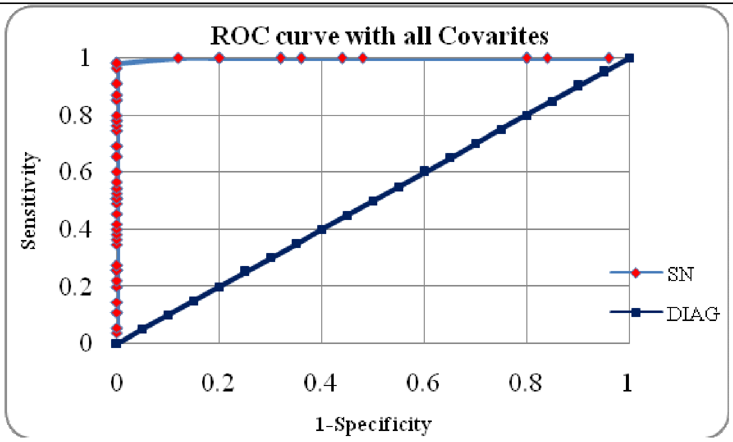

Figure: 5 ROC curve with all Covariates